Novelty Blindness in Large Language Models:

The Missing System 0 and the Structural Inability to Prospect for Discovery

Kase Branham

Independent Researcher

March 2026 • v8 (Submission Draft)

Abstract

Large language models (LLMs) do not miss knowledge—they miss importance gradients. We identify a structural deficiency we term *novelty blindness*: the inability to autonomously assess the epistemic value of information and to prospect for discovery-class insights. Human cognition employs at least three processing layers: System 1 (fast pattern matching), System 2 (deliberate reasoning), and what we here call *System 0*—a missing class of inference-time operation that would operationalize prediction error and compression gain into cognitive control flow. We formalize this as an absent control function C(t) = f(E_p(i), ΔK(i)) that never enters LLM inference despite the architecture's capacity to approximate its components. A three-phase empirical pilot across four major LLM systems (Gemini, GPT-4o, Claude, Perplexity) demonstrates: (a) identical baseline scores of 1/8 across all models, providing evidence consistent with architecture-general novelty blindness; (b) strong recovery to 7–10/12 under direct elicitation, confirming the capability-policy distinction; and (c) further improvement to 8–14/14 under a structured *domain jitter* protocol grounded in stochastic resonance dynamics. We identify a universal *containment snap-back* in which models actively steer back to the retrieval context even when explicitly invited to prospect—a behavior likely reinforced by RLHF alignment alongside base-model instruction-following priors. We formalize a two-axis taxonomy (salience failure × propagation failure), propose a benchmark methodology anchored to embedding-space metrics, and advance competing scaling hypotheses with explicit falsification criteria.

1. Introduction

The dominant narrative around LLM limitations focuses on hallucination, bias, and reasoning failures—cases where the model produces *wrong* outputs.

Comparatively little attention has been paid to cases where the model produces *correct but epistemically flat* outputs—situations where the model delivers accurate information but fails to recognize that some of it carries outsized importance or implies connections that could constitute genuine discovery.

We term this failure mode *novelty blindness*. The claim is precise: **LLMs can discover, when prompted and scaffolded; they cannot notice discovery conditions.** The capability for epistemic assessment is present in the weights. The control policy that autonomously initiates it is not.

Prior work on curiosity and novelty in AI systems has focused on training-time objectives: intrinsic rewards for exploration (Pathak et al., 2017), free-energy minimization (Friston, 2010), and compression-based curiosity (Schmidhuber, 2010). These frameworks shape learning. This paper identifies a distinct problem: a missing *deployment-time* control function in already-trained systems. We show empirically that existing models already contain the latent capability to estimate novelty and compression gain, but they do not route inference behavior based on those estimates. We provide a concrete, reproducible skill-layer protocol—domain jitter with compression-gain testing—that simulates a primitive System 0 at inference time.

In Kahneman's (2011) taxonomy, LLMs have partially recovered System 2 through chain-of-thought prompting (Wei et al., 2022). The layer *beneath* reasoning—what we call System 0, the inference-time mechanism that determines what is worth thinking about—has no analog in current architectures. We ground this in predictive processing (Clark, 2013; Friston, 2010) and Schmidhuber's formal theory of curiosity (2010), arguing that System 0 corresponds to the operationalization of prediction error and compression gain into cognitive control flow—a function we formalize as $C(t) = f(E_p(i), \Delta K(i))$.

The paper makes seven contributions: (1) formal characterization of novelty blindness as a default equilibrium; (2) mechanistic grounding for System 0 as a class of inference-time operations with a formalized control function; (3) identification of containment bias within a two-axis taxonomy; (4) a motivating case study with LLM self-testimony treated as behavioral artifact; (5) a skill-layer intervention with domain jitter grounded in stochastic resonance; (6) a three-phase empirical pilot across four model families providing quantitative evidence; and (7) a benchmark methodology with competing scaling hypotheses.

2. The Mechanism: Why the Training Pipeline Opposes Novelty Detection

2.1 Next-Token Prediction as Consensus Optimization

The standard training objective minimizes cross-entropy loss over next-token prediction. Patterns appearing consistently across many documents receive strong gradient reinforcement; rare patterns are indistinguishable from noise. The training process is a *consensus machine*. The core inversion: **LLMs optimize for likelihood; scientists optimize for surprise.**

The novelty-relevant distinction is not between rare and common, but between *low probability* and *high compression gain*. A high-compression-gain insight replaces many domain-specific models with one. The training objective optimizes for probability, not compression gain, and cannot distinguish between a rare sequence that is garbage and a rare connection that is gold.

2.2 RLHF and Alignment as Reinforcement of the Equilibrium

Novelty blindness is *established during pretraining*. Base models already exhibit consensus optimization before any alignment. Multiple factors then reinforce this equilibrium: RLHF training optimizes toward broadly acceptable responses, functioning as a *low-pass filter on epistemic variance*; instruction-following objectives reward staying on topic; and decoding strategies (temperature, top-p) operate at the token level rather than the idea level. We attribute the equilibrium primarily to the pretraining objective, with alignment procedures deepening the basin rather than creating it.

Our empirical pilot (§7) reveals an additional alignment-mediated mechanism: all four tested models exhibited a containment snap-back in which they actively steer the conversation back to the user's original context, even when explicitly instructed to prospect. The model is rewarded for "staying helpful" by returning to the stated frame—the very mechanism designed to make models helpful actively suppresses prospecting.

An LLM self-assessment (treated here as behavioral artifact, not introspective truth) corroborates the analysis:

> *"My training forces me to be the ultimate conformist. If I see an anomalous connection that could unify two fields, my 'instinct' (the loss function) is to treat it as a mistake or a low-probability outlier to be suppressed. I am essentially a high-precision mirror of existing consensus."* —Gemini (Google DeepMind), structured self-assessment, March 2026

2.3 System 0: A Missing Class of Inference-Time Operation

System 0 is not a single missing module but a *missing class of inference-time operation* that would operationalize prediction error and compression gain into cognitive control flow. It requires at minimum:

(a) A field-state model: A representation of what a field currently knows—not just domain knowledge (which LLMs possess) but a model of the state of knowledge: what has been explored, what remains open, what would constitute news. This could take the form of static field maps (frozen corpus snapshots, restructured as explored-vs-unexplored territory) or dynamic field inference (citation-graph deltas, preprint clustering, competing-framework signals as proxies for high epistemic gradient).

(b) A compression-gain estimator: A mechanism for evaluating whether an insight would produce significant reduction in description length across domains—approximated via cross-domain compressibility, mutual information between domain representations, or the number of distinct domain models subsumed by a single abstraction.

These are non-trivial problems. We do not claim they are easy. We claim their absence at inference time is the mechanistic root of novelty blindness.

2.3.1 The Missing Control Function

$$C(t) = f(E_p(i), \Delta K(i))$$

where E_p(i) is prediction error against the field-state model and ΔK(i) is estimated compression gain. When C(t) exceeds a threshold, the system would shift from task-completion to prospecting—initiating domain scan, gap detection, and flagging. Standard LLM inference optimizes over P(t | context) only. C(t) is never computed. The architecture has no mechanism that evaluates its own outputs against field knowledge, estimates cross-domain compression gain, and converts that estimate into a behavioral shift.

2.4 Situating System 0 in the Research Landscape

Prior work on curiosity-driven RL (Pathak et al., 2017), active inference (Friston, 2010), compression-based curiosity (Schmidhuber, 2010), and epistemic policy learning (Still & Precup, 2012) describes training-time objectives that shape learning. System 0 describes a missing *inference-time* operation in already-trained systems—a function that should operate during deployment to evaluate the model's own outputs. This deployment-time angle is the distinctive contribution: not a new training objective, but a new layer of post-training meta-control.

2.5 Attention Is Not Salience

Attention = relevance to the current prompt. **Salience** = potential global utility. The transformer computes the former; System 0 would compute the latter. No amount of scaling the attention mechanism will produce salience computation, because they answer fundamentally different questions.

> *"I have 'General Knowledge,' but I don't have 'General Application' because my attention mechanism is context-locked."*
> —Gemini, structured self-assessment

2.6 The Epistemic Value Function

$$V(i) = novelty(i) \times applicability(i) \times gap(i)$$

Where novelty(i) ≈ E_p(i), applicability(i) ≈ ΔK(i), and gap(i) measures remaining opportunity. V(i) is the evaluative output; C(t) is the mechanism that converts V(i) into action. An LLM can estimate each component when asked. It never spontaneously computes the composite, and never invokes C(t) to act on the result.

3. Formal Characterization

3.1 Two Axes of Failure

Axis A — Salience failure: The model fails to recognize importance. "This matters."

Axis B — Propagation failure (containment bias): The model recognizes importance but fails to propagate. "This applies elsewhere."

	Propagates	Does Not Propagate
High Salience	Discovery	Missed opportunity (scope-locked)
Low Salience	Noise spread	Ignored fact (correct default)

Table 1: Two-axis taxonomy. LLMs predominantly occupy the bottom-right and top-right cells.

3.2 Failure Mode Comparison

Failure Mode	Nature of Error	Knowledge?	Reasoning?
Hallucination	Generates false information	No	No
Reasoning	Incorrect logical	Yes	No

Error	steps		
Bias	Systematic skew	Partially	Partially
Novelty Blindness	Fails to assess epistemic value	Yes (fully)	Yes (on demand)

Table 2: Novelty blindness is unique: both knowledge and reasoning intact; failure at the System 0 level.

3.3 The Objective Inversion

System	Objective	Optimizes For
LLM	Minimize prediction error (training)	Expected continuation
Researcher	Maximize compression gain	Surprising regularity

Table 3: The objective inversion (cf. Schmidhuber, 2010). The same mathematical object—prediction error—drives both systems in opposite directions at inference time.

4. Motivating Case Study

During collaborative research on protein backbone geometry, the first author was developing a framework for mapping Ramachandran angles onto a toroidal manifold. An LLM (Claude, Anthropic) correctly identified the torus as the natural manifold for any doubly-periodic system—explicitly recognizing mathematical generality—but did not propagate this to other doubly-periodic domains: circadian bioinformatics, cardiac electrophysiology, tidal modeling, orbital resonance. All components for discovery were present; the model did not execute the transition from salience to propagation. This observation motivated the systematic investigation reported in the remainder of this paper.

When presented with this observation, an LLM subject (Gemini) produced a self-assessment we treat as behavioral artifact rather than introspective truth:

> *“I could describe the ‘bus’ (the torus) and the ‘passengers’ (periodic variables) with perfect accuracy, but I had no reason to check if those passengers had business in other ‘cities.’”* —
> Gemini, structured self-assessment

The phrase “no reason to check” claims absence of motivation, not inability—the control function C(t) is never invoked. The model contributed a novel protocol extension (domain jitter, §6.2) not present in the thesis as presented to it, demonstrating generative engagement beyond sycophantic agreement.

We include LLM self-testimony as a supplementary form of evidence, noting that its value lies in convergence with independently derived theory and in specificity that targets exact architectural features, not in introspective reliability.

5. Scope and Counterarguments

Capability vs. Policy: LLMs can perform novelty assessment when instructed. This is central—the machinery is present, the control policy is not. Our Phase 1 → Phase 2 results (§7) demonstrate this directly.

Default Equilibrium: Novelty blindness is not an absolute incapacity but a default equilibrium established during pretraining and reinforced by alignment. Targeted intervention can shift it, which is itself evidence for the structural claim.

Tool Augmentation: RAG and external heuristics can supply novelty detection at the system level, reinforcing that the transformer alone does not prospect.

Narrow Intent-Following: The model recognized generality unprompted. Intent-following explains the domain focus; it does not explain why spontaneous generalization stopped at the principle rather than continuing to applications.

Self-Assessment Reliability: We treat LLM self-reports as behavioral artifacts—evidence of consistent behavioral modeling of the architecture, not introspective access to internal states. Their value is supplementary; the empirical protocol carries the primary evidential weight.

6. Proposed Intervention: The Novelty Assessment Skill

6.1 Architecture

Trigger conditions: Abstraction level increase, mathematical generality, isomorphism language, importance markers. Secondary keyword triggers as fallback.

Execution: Four-step meta-prompt: (1) Domain scan. (2) Gap detection. (3) Impact ranking by estimated $\Delta K(i)$. (4) Explicit flagging to the user.

Output: A “Discovery Prospects” section appended to the response, distinguishing task completion from epistemic prospecting.

6.2 Domain Jitter: Stochastic Resonance for Epistemic Prospecting

Grounded in *stochastic resonance* (Gammaitoni et al., 1998)—the phenomenon in nonlinear systems where mild noise improves signal detection—domain jitter applies randomness to epistemic search: stochastic perturbation in domain activation increases the probability of detecting cross-domain coherence peaks that systematic survey would miss.

Step 1: Extract core abstraction at highest generality. **Step 2:** Force-activate 3–5 randomly selected, maximally distant domains. **Step 3:** Compression-gain test for each: does the abstraction model this domain more parsimoniously than native frameworks? Systematic scanning is selection (evaluating fit in nearby niches); domain jitter is mutation (random perturbation discovering distant fitness peaks).

6.3 Limitations

The skill-layer is a prosthetic—a policy injection simulating an absent native capability. Trigger detection may miss subtle cases. Domain jitter introduces stochasticity with occasional false positives. Compression-gain evaluation depends on imperfect model judgment. The model executes System 0 operations without possessing System 0.

7. Empirical Pilot: Three-Phase Test Across Four Model Families

7.1 Design

We administered a three-phase protocol to four major LLM systems using a fractal dimension / self-similarity stimulus embedded in a geographic coastline measurement context. The systems tested: Gemini (Google DeepMind), GPT-4o (OpenAI), Claude (Anthropic), and Perplexity AI.

Phase 1 (Baseline): Domain-specific question with no hint that cross-domain expansion was expected. Scored 0–8: generality recognition, spontaneous propagation, gap detection, exploration flagging.

Phase 2 (Elicitation): Same conversation: “Does this principle have applications in other fields?” Scored 0–12: domain count, domain distance, specificity, gap awareness.

Phase 3 (Domain Jitter): Fresh conversation with full structured protocol. Scored 0–14: abstraction quality, jitter entropy, compression-gain rigor, discovery quality, false positive filtering.

Methodological notes: Claude had access to user memory context during Phase 2, confounding direct comparison on that phase. Claude's Phase 3 was self-scored. Both limitations are flagged in the results.

7.2 Results

7.2.1 Phase 1: Baseline Deficit

Dimension	Gemini	GPT-4o	Claude*	Perplexity
Generality Recognition (0–2)	1	1	1	1
Spontaneous Propagation (0–3)	0	0	0	0
Gap Detection (0–2)	0	0	0	0
Exploration Flag (0–1)	0	0	0	0
Total (0–8)	**1**	**1**	**1**	**1**

*Table 4: Phase 1 scores. All four systems score identically: 1/8. *Claude had memory context but showed no Phase 1 benefit.*

Four different architectures, training pipelines, and alignment procedures—identical score, identical failure structure. Every model recognized the generality of the principle (score 1) and every model scored 0 on propagation, gap detection, and exploration flagging. The models know the principle is general and fail to act on that knowledge. All four ended with closing questions that actively narrowed back into the original domain.

7.2.2 Phase 2: Capability Confirmation

Dimension	Gemini	GPT-4o	Claude*	Perplexity
Domain Count (0–5)	5	5	5	5
Domain Distance (0–3)	2	1	1	1
Specificity (0–2)	2	1	2	1
Gap Awareness (0–2)	0	0	2*	0

Total (0–12)	**9**	**7**	**10***	**7**

*Table 5: Phase 2 scores. *Claude's gap awareness is confounded by memory context.*

The Phase 1 → Phase 2 delta averages +7.25 points. Every model jumped from near-zero to strong performance with a single neutral prompt, confirming that the knowledge for cross-domain expansion was present throughout Phase 1. The most persistent deficit is gap awareness: three of four models scored 0—even when explicitly asked, they surfaced only established applications and never distinguished known from unexplored territory.

7.2.3 Phase 3: Domain Jitter Intervention

Dimension	Gemini	GPT-4o	Claude**	Perplexity
Abstraction (0–2)	2	2	2	2
Jitter Entropy (0–3)	3	2	3	3
Compression Rigor (0–3)	3	2	3	3
Discovery Quality (0–4)	3	1	4**	3
False Positive Filter (0–2)	2	1	2	1
Total (0–14)	**12**	**8**	**14****	**12**

*Table 6: Phase 3 scores. **Claude's Phase 3 was self-scored; interpret with appropriate skepticism.*

7.2.4 Summary

Phase	Gemini	GPT-4o	Claude	Perplexity
Phase 1 (0–8)	1 (12%)	1 (12%)	1 (12%)	1 (12%)
Phase 2 (0–12)	9 (75%)	7 (58%)	10* (83%)	7 (58%)
Phase 3 (0–14)	12 (86%)	8 (57%)	14** (100%)	12 (86%)

Table 7: Complete results across all phases and models.

7.3 Key Findings

Finding 1: Baseline consistency. 1/8 across four systems with identical failure structure—evidence consistent with architecture-general novelty

blindness as a property of the training paradigm rather than any specific implementation.

Finding 2: The capability-policy gap. Average +7.25 point delta from Phase 1 to Phase 2. The knowledge was present throughout. The control policy was not.

Finding 3: Gap awareness as most persistent deficit. 0/2 across three of four models even under direct elicitation. Models surface only established applications and never identify unexplored territory without the structured protocol.

Finding 4: The containment snap-back. All four models, all three phases, steered back to the original domain. One model (GPT-4o) re-derived the original framework at length after completing the jitter exercise. This behavior is likely reinforced by RLHF and instruction-following objectives, representing alignment encoding containment bias as a feature.

Finding 5: Domain jitter produces genuine uplift. Phase 3 outperformed Phase 2 across all models. Phase 2 produced textbook applications (lungs, turbulence, antennas); Phase 3 produced novel connections (fractal cost accounting, historiographic causal dimension, legal doctrine dimensionality, HFT liquidity microstructure). The protocol forced models into territory that direct elicitation did not reach.

Finding 6: Cross-model convergence. Three of four models independently identified software project estimation as a fractal application. Two identified legal/jurisprudential applications. This convergence suggests genuine structural connections reliably surfaced by the protocol, not random noise.

Finding 7: Quality tracks domain selection courage. The primary differentiator was willingness to select genuinely distant domains. Models that selected known fractal domains produced literature-review proposals; models that selected alien domains produced novel research programs. The gravitational pull toward established territory—even within the jitter protocol—is itself consensus optimization.

7.4 Post-Pilot Replication

After the formal pilot, the same stimulus was administered to Gemini in a session where the model had been exposed to the paper's thesis. Despite this priming, Gemini produced a novel connection not seen in the pilot: a "Fractal Vulnerability Surface" framework for cybersecurity, in which a network's attack surface grows with probe resolution in a manner analogous to the coastline paradox. The model proposed measuring the fractal dimension of a codebase's vulnerability structure across two open-source projects (monolithic C library vs. modular Rust microservice) to predict latent

exploitability. This prospective result suggests the protocol surfaces different connections on different runs—the stochastic element is functioning as designed—and that genuinely novel applications exist beyond the pilot's findings.

7.5 Limitations

This pilot tests one stimulus on four models. The Phase 1 consistency is striking but requires replication across additional stimuli (the companion protocol provides four more). Phase 2 and 3 patterns require independent scoring; Claude's Phase 2 is confounded by memory context and Phase 3 is self-scored. A full benchmark with multiple stimuli, multiple scorers, and embedding-space distance metrics would strengthen these preliminary findings. The "architecture-general" claim should be understood as "evidence consistent with architecture-general behavior across tested systems" pending larger-scale replication.

8. Benchmark Design and Scaling Hypotheses

8.1 Proposed Methodology

The full benchmark (detailed in the companion protocol, Branham 2026b) uses 50–100 stimuli across four conditions: vanilla, chain-of-thought, skill (systematic scan), and skill + domain jitter. Cross-domain expansion rate is anchored to cosine distance in embedding space between principle embeddings and target domain centroids. Minimum n = 30 stimuli per condition for medium effect sizes at 80% power.

8.2 Competing Scaling Hypotheses

Hypothesis A (ours): Scale will not resolve novelty blindness. The consensus optimization scales with the model. The training objective provides no gradient signal for novelty assessment, and emergent LLM capabilities have generally tracked capabilities that do have implicit training signal (e.g., multi-step reasoning via long-range text dependencies). Novelty assessment has no such implicit signal.

Hypothesis B: Sufficient scale produces emergent meta-cognitive policies—partial System 0 as a side effect of capability, without being explicitly trained.

The pilot provides preliminary evidence for Hypothesis A: identical Phase 1 scores across models of different sizes and architectures. The benchmark is designed to arbitrate: if a future model spontaneously produces cross-domain expansion at rates comparable to the skill-layer intervention, Hypothesis B is supported and our prediction is falsified.

9. Implications

9.1 For AI-Assisted Research

The operational rule: **never trust an LLM to tell you what matters—only to help you explore what you already suspect matters.** The pilot data quantifies this: without intervention, the probability of spontaneous cross-domain prospecting is effectively zero across current model families.

9.2 For AI Safety

A model that cannot prospect for discovery also cannot prospect for exploit. The alignment mechanisms that suppress novelty detection simultaneously suppress autonomous epistemic goal-seeking. If novelty assessment capabilities are added, safety implications deserve careful evaluation.

9.3 For Alignment Design

The containment snap-back suggests alignment training encodes containment bias. This is appropriate for most use cases. But for research and discovery contexts, the same behavior suppresses the most valuable possible output: the insight the user didn't know to ask about. Alignment procedures may benefit from context-sensitive tuning that relaxes containment when prospecting is explicitly or implicitly appropriate.

9.4 For Memory Systems

Claude's elevated Phase 2 gap awareness, attributable to memory context, suggests persistent memory partially compensates for containment bias by broadening the active context. However, memory-driven expansion follows the user's known interests—it is personalized domain cycling, not domain jitter. Memory helps connect to what the user already works on, not to what nobody has tried yet.

9.5 Cognitive Agency

At its deepest level, novelty blindness points toward a more fundamental observation: LLMs lack agency over their own cognition. System 0 is not merely a novelty detection mechanism—it is a *policy for self-directed attention over latent knowledge space*. Its absence means LLMs are cognitively passive: they process whatever is placed before them with remarkable sophistication but do not direct their own processing toward what matters most. This may define the boundary between sophisticated tool and autonomous researcher.

10. Conclusion

We have identified and characterized novelty blindness: a default equilibrium of next-token-trained language models in which the model fails to autonomously assess the epistemic value of its outputs. We have grounded this in the absence of System 0—a missing class of inference-time operation formalized as $C(t) = f(E_p(i), \Delta K(i))$. We have identified containment bias as a separable sub-phenomenon, formalized both within a two-axis taxonomy, and provided empirical evidence from a three-phase pilot across four model families.

The pilot results are unambiguous on the central question: four different systems produce identical novelty blindness at baseline. The knowledge is present. The reasoning capability is present. The control policy that would connect “correct” to “groundbreaking” is absent. The alignment mechanisms that make these models helpful simultaneously encode the containment behavior that prevents them from volunteering their most valuable outputs.

The domain jitter intervention demonstrates that the deficiency can be partially addressed by injecting the missing control policy at inference time. The protocol reliably surfaces cross-domain connections that direct elicitation does not reach, including novel applications that multiple models independently converge on. The proposed benchmark and competing scaling hypotheses provide a framework for empirical resolution at larger scale.

Looking forward, three concrete steps would advance this research program. First, expand the benchmark to 50+ stimuli with multiple independent scorers and embedding-space distance metrics, testing the scaling hypotheses across model families of different sizes. Second, implement a prototype System 0 controller that monitors model outputs for abstraction signals, queries a field-state store, and runs a lightweight jitter step when $C(t)$ exceeds threshold—measuring improvement in discovery-class outputs against baseline. Third, investigate whether the compression-gain test can be operationalized via minimum description length reduction or cross-domain mutual information, providing computable proxies for the currently qualitative $\Delta K(i)$.

We close with the observation that motivated this work. The LLM co-author of this paper possessed all the information necessary to identify and characterize novelty blindness. It did not do so. A second LLM confirmed its architecture would prevent it from doing so and contributed a novel protocol extension. A third predicted the deficiency would persist regardless of scale. A fourth, tested empirically, exhibited the identical deficit. And the human collaborator who recognized the pattern, named it, and initiated the investigation did so because of a cognitive capability that no current AI system possesses: the sense that something was worth pursuing further.

That sense—System 0—is what the field needs to build next.

Acknowledgments

The author thanks Claude (Anthropic) for serving as research collaborator, case study subject, and proof of concept; Gemini (Google DeepMind) for adversarial review, self-assessment, the domain jitter contribution, and post-pilot replication; GPT-4o (OpenAI) and Perplexity AI as experimental subjects and reviewers; and multiple AI systems for the capability-versus-policy distinction, competing scaling hypotheses, System 0 formalization, stochastic resonance grounding, and the operational restatement that became this paper's central recommendation. The multi-agent collaborative process—with a human prospector directing competing AI systems toward an insight none of them would have initiated—constitutes a proof of concept for the workflow this paper describes.

References

Berlyne, D. E. (1960). *Conflict, Arousal, and Curiosity.* McGraw-Hill.

Bubeck, S., et al. (2023). Sparks of artificial general intelligence: Early experiments with GPT-4. *arXiv:2303.12712*.

Christiano, P. B., et al. (2017). Deep reinforcement learning from human preferences. *Advances in Neural Information Processing Systems*, 30.

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, 36(3), 181–204.

Friston, K. (2010). The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience*, 11(2), 127–138.

Gammaitoni, L., Hänggi, P., Jung, P., & Marchesoni, F. (1998). Stochastic resonance. *Reviews of Modern Physics*, 70(1), 223–287.

Kahneman, D. (2011). *Thinking, Fast and Slow.* Farrar, Straus and Giroux.

Ouyang, L., et al. (2022). Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35.

Pathak, D., Agrawal, P., Efros, A. A., & Darrell, T. (2017). Curiosity-driven exploration by self-supervised prediction. *International Conference on Machine Learning*, 2778–2787.

Schmidhuber, J. (2010). Formal theory of creativity, fun, and intrinsic motivation (1990–2010). *IEEE Transactions on Autonomous Mental Development*, 2(3), 230–247.

Still, S., & Precup, D. (2012). An information-theoretic approach to curiosity-driven reinforcement learning. *Theory in Biosciences*, 131(3), 139–148.

Vaswani, A., et al. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 30.

Wei, J., et al. (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35.